AD-A050 247 NAVAL POSTGRADUATE SCHOOL MONTEREY CALIF F/G 12/1
ANALYSIS AND MODELLING OF POINT PROCESSES IN COMPUTER SCIENCE.(U)
SEP 77 P A LEWIS, G S SHEDLER
UNCLASSIFIED NPS55-77-38 NL

1 OF 1

AD
A050 247

END
DATE
FILMED
3-78
DDC

2

AD A050247

AD No.
DDC FILE COPY

NPS55-77-38

NAVAL POSTGRADUATE SCHOOL

Monterey, California

ANALYSIS AND MODELLING OF POINT PROCESSES

IN COMPUTER SYSTEMS

by

P. A. W. Lewis and G. S. Shedler

September 1977

Approved for public release; distribution unlimited.

D D C
RECEIVED
FEB 22 1978
D

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

Rear Admiral Isham Linder
Superintendent

Jack R. Borsting
Provost

Reproduction of all or part of this report is authorized.

This report was prepared by:

P.A.W. Lewis
P. A. W. Lewis, Professor
Department of Operations Research

G. S. Shedler (by P.A.W.L.)
Gerald S. Shedler
IBM Research Laboratory

Reviewed by

Michael G. Sovereign
Michael G. Sovereign, Chairman
Department of Operations Research

Robert R. Fossum
Robert Fossum
Dean of Research

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

(14) 1. REPORT NUMBER
NPS55-77-38

2. GOVT ACCESSION NO.

(9) 3. RECIPIENT'S CATALOG NUMBER

(6) 4. TITLE (and Subtitle)
Analysis and Modelling of Point Processes in Computer Science,

5. TYPE OF REPORT & PERIOD COVERED
Technical rept.,

6. PERFORMING ORG. REPORT NUMBER

(10) 7. AUTHOR(s)
P. A. W./Lewis [redacted] G. S./Shedler

8. CONTRACT OR GRANT NUMBER(s)
ONR Grant NR-42-284
NSF ~~Grant AF476~~

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Naval Postgraduate School
Monterey, Ca. 93940

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Naval Postgraduate School
Monterey, Ca. 93940

(11) 12. REPORT DATE
Sep[redacted] [redacted]77

13. NUMBER OF PAGES
44 (12) 45p.

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Univariate and Multivariate Series of Events, Statistical Analysis of Point Processes, Spectral Analysis, Trend Analysis, Positive Multivariate Time Series, Computer System Models, Queueing Networks, Stochastic Point Processes Computer Reliability, Cluster Processes, Page Exception Processes, CONT.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Models of univariate and multivariate series of events (point processes) and statistical methods for the analysis of point processes have diverse applications in the study of computer systems. ~~We review~~ these applications, which include the analysis and prediction of computer system reliability and the evaluation of computer system performance, are reviewed with emphasis on the latter. In addition ~~we describe~~ recent results are described in the development of methodology for the statistical analysis of point processes. ~~We point out that~~ the analysis of multivariate point processes is much more difficult than that of

next page

251 450

19. Keywords cont.
Nonhomogeneous Poisson Process, Multiprogramming, Multiprogrammed Computer Systems, Autoregressive Sequences, Cluster Process, Discrete Time Series, Self-Similar Processes.

20. Abstract cont.
univariate point processes, and that methodology has only recently been developed in a perforce fairly tentative manner. The applications to computer system data illustrate the need for new data analytic methods for handling large amounts of data, and the need for simple models for non-normal, positive multivariate time series. Some starts in these directions are indicated.

ANALYSIS AND MODELLING OF POINT PROCESSES IN COMPUTER SYSTEMS

P. A. W. Lewis and G. S. Shedler

Naval Postgraduate School, Monterey, California 93940
IBM Research Laboratory, San Jose, California 95193

ABSTRACT

Models of univariate and multivariate series of events (point processes) and statistical methods for the analysis of point processes have diverse applications in the study of computer systems. We review these applications, which include the analysis and prediction of computer system reliability and the evaluation of computer system performance, with emphasis on the latter. In addition we describe recent results in the development of methodology for the statistical analysis of point processes. We point out that the analysis of multivariate point processes is much more difficult than that of univariate point processes, and that methodology has only recently been developed in a perforce fairly tentative manner. The applications to computer system data illustrate the need for new data analytic methods for handling large amounts of data, and the need for simple models for non-normal, positive multivariate time series. Some starts in these directions are indicated.

ACCESSION for
NTIS White Section ☒
DDC Buff Section □
UNANNOUNCED □
JUSTIFICATION ....
BY ....
DISTRIBUTION/AVAILABILITY CODES
Dist. AVAIL. and/or SPECIAL
A

DDC
RECEIVED
FEB 22 1978
D

ANALYSIS AND MODELLING OF POINT PROCESSES IN COMPUTER SYSTEMS

P. A. W. Lewis and G. S. Shedler

Naval Postgraduate School, Monterey, California 93940
IBM Research Laboratory, San Jose, California 95193

1. INTRODUCTION

In this paper we review applications of models for univariate and multivariate series of events (point processes) in computer systems and statistical methods for the analysis of these point processes. There are many examples of such series of events. Typical examples include the following:

(i) occurrences of system failure. These events may be typed as "hardware" failures or "software" failures, giving a bivariate point process. They may also be typed by the physical part of the system in which the failure occurred;

(ii) arrivals of requests to a storage subsystem. These events may be marked by an identifier of the requested record;

(iii) occurrences of exceptions in a system having hierarchical storage. These events may be typed according to the level of the hierarchy at which required information is found.

The applications to computer system problems of point processes methodology have been to computer system reliability and to computer system performance evaluation. In Section 2 we review these applications, with emphasis on the very broad area of performance

September 5, 1977

evaluation. What emerges is the need for new data analytic methods for the particular problems encountered, and the need for simple models for positive multivariate time series when the data analysis is used to suggest models or modify postulated models. Some recent results in the development of methodology for statistical analysis of point processes are described in Section 3. A final section indicates some starts on the development of some new and particularly suitable models for non-normal time series.

2. POINT PROCESSES OCCURRING IN COMPUTER SYSTEMS

2.1. Computer System Reliability

We describe first an application of point process methodology in computer system reliability studies, namely the analysis and modelling of computer failure patterns.

2.1.1. Computer system failure patterns. The earliest application of point process methodology to problems associated with computer systems is the analysis and modelling of computer failure patterns given by Lewis (1964a). A primary motivation for the construction of computer failure models is to analyze data from operational systems and to find ways of comparing and perhaps improving the reliability of existing systems. Prediction and optimization of the reliability of future systems is a second motivation.

Reliability models for complex systems which were current in 1964 predict that the failure pattern of a computer system should form a Poisson process. This model is derived from the assumption that the failures in each component position constitute independent renewal processes. The failure pattern of the system is then formed from the pooled failures of the components. Under the stochastic assumptions that are made, it is known (Cox and Smith, 1954) that the pooled series of events will be indistinguishable from a Poisson process over periods of time which are short compared with the mean times-to-failure of the components. The data presented by Lewis (1964a) show, however, that the times-between-failures of large computer systems are not exponentially distributed and that successive times-between-failures are correlated. Physically the observed clustering of failures and the resulting departures from a Poisson process arise from imperfect repair, i.e., because failed components are not always located and removed the first time they cause system failure, nor are the failed components always needed for correct system operation. Subsidiary system failures are then induced a short time later.

The Lewis (1964a) paper deals with the development of a model (the branching Poisson process) for computer failure patterns which accounts for the observed departures from a Poisson process. The probabilistic properties of the model are derived and used to analyse three series of computer failures (consisting of 109, 186, and 255 events respectively). Lewis (1964b) discusses implications of the model for the use and maintenance of computer systems.

We describe the branching Poisson process model briefly; in the model times between original failures of components constitute a main process $\{X_i\}$. At each point of this process an attempt is made to locate and repair the failure, the attempt succeeding with fixed probability, independently of other attempts to repair main failures. Otherwise the failure recurs at times Y_1, $Y_1 + Y_2$, ... , $Y_1 + \cdots + Y_{S+1}$ after the initial occurrence. Thus, S+1 unsuccessful attempts are made in all to locate and remove the source of the computer failure. The computer failure pattern is then the superposition of the events in the main process and the events in the subsidiary processes which the main events generate. The $\{X_i\}$ are assumed to be independent, identically and exponentially distributed and the intervals Y_i in the subsidiary processes are assumed to be mutually independent and identically distributed. The branching Poisson process model is also called the Barlett-Lewis cluster process or a Poisson cluster process. Properties of the model have been developed by several authors; see e.g., Lewis (1969), Oakes (1975).

The result of the statistical analysis given by Lewis (1964a) is the demonstration that the differing rates of failure in the three systems under study is due to the adequacy or inadequacy of the maintenance on each system. This was done by estimating (rather crudely) the rate of failures in the main Poisson process and the expected number, E(S+1), of unsuccessful attempts to fix an original failure. Then E(S+1) measures the adequacy of the maintenance of the system. We know of no advance in the statistical analysis of Poisson cluster processes since the Lewis (1964a) paper.

2.2. Computer System Performance Evaluation

Because of the complexity of existing and proposed computer systems, detailed measurements of running systems are needed in order to develop system models. This measurement and modelling comprises just one facet of computer system performance evaluation. Ultimate goals of performance evaluation include tuning of existing systems and prediction (usually via simulations) of the performance of proposed systems. For example, it is desirable to have an airline reservation system which is efficient from both the customers' and airline's points of view in the sense that it should respond quickly and reliably at a reasonable cost.

Given the complexity of computer systems and the resulting relative difficulty of carrying out meaningful performance evaluations and designs, the collection and analysis of measurement data from representative systems to identify and characterize significant performance phenomena is necessary. The availability of such measurements presents the possibility of obtaining thereby empirically valid, parameterized mathematical models for the workload of the system. For performance evaluation studies, in addition to workload or program behavior models, the analyst needs a model for the computer system or subsystem structure and frequently uses a network of queues as a system model. Such networks provide a convenient means of representing the interaction between the processing and input-output resources of (multiprogrammed) computer systems and subsystems. There is a large literature dealing with queueing network models; see e.g., Gaver (1967), Lewis and Shedler (1971), Buzen

(1971), Moore (1971), and Lavenberg and Shedler (1976). Under the usual convenient, but not necessarily realistic, queueing-theoretic assumptions (e.g., independent and identically often exponentially, distributed service times) analyses of queueing network models based on a "numbers-in-queue" state space can be carried out; see Jackson (1963), Baskett, Chandy, Muntz and Palacios (1975), Reiser and Kobayashi (1975), Kelly (1975), (1976) and Gelenbe and Muntz (1976). These analyses yield expressions for stationary queue length distributions that can be evaluated numerically to provide measures of system performance such as device "utilizations" and job "throughput." All these models use global assumptions of independent; measurements usually indicate, for example, that interarrival times are correlated and incorporating this dependence in the model will sometimes give very different results; see, for example, Jacobs (1977).

Other measures of system performance (calculated as sums of queueing times) involve the distribution of times for a job to traverse a portion of the network. These times (in closed networks complete circuits or loops, and in open networks times from source to sink) are often interpretable as job "response times" and these response times are likely to be particularly sensitive to workload characteristics. Analyses based on the numbers-in-queue state space yield expected values for response times, but do not yield other characteristics of interest such as percentiles or quantiles. Since alternative analyses to provide these characteristics are in general not available, it is necessary to undertake simulation studies of the queueing networks. Such simulations, and indeed simultations of more complex queueing networks under more realistic stochastic assumptions, are inherently difficult, and are likely to be time-consuming and costly to carry out.

We describe three applications of point process methodology in computer system performance evaluation: analysis of page exceptions in a two-level memory, analysis of exceptions in a three-level storage hierarchy, and analysis of transaction processing in a data base management system.

2.2.1. Page exceptions in a two-level memory. The following brief description provides the computer system context for the discussion of page exception processes given below. We consider a system whose memory resource (for storage of information) comprises a main memory and an auxiliary memory, and assume that main memory is the execution store, i.e., only information that is resident therein can be processed. We also assume that the auxiliary memory is large enough to hold all information required by a program which is to be processed. When such a system operates in a so-called paging environment, units of equal size called pages partition all of the information that is explicitly addressable by the single processor (central processing unit). Similarly, page-size sections called page-frames partition main memory. It is possible to execute a program by supplying it with only a few page-frames of main memory, as follows. When the page containing the first executable instruction has been loaded into some page frame, execution begins and continues until the program requires some information not found in main memory. The operating system fetches the page containing the missing information from auxiliary memory (overwriting some page currently in main memory), and execution of the program continues, and so forth. In demand paging, information enters main memory only as a result of an attempt (detected by the system hardware) to use information not currently in main memory. A page exception is an instance of this implicit "demand" for a page which is not in main memory. When dealing with large programs or in a multiprogramming mode in which main memory is shared among several programs it is usually the case that main memory is filled when the system must fetch another page from auxiliary memory. Consequently, it is necessary to choose a page frame in main memory to be overwritten. The replacement algorithm is the rule governing this choice. Most of the time, before overwriting the chosen page frame, the system must save the content of the page frame. (See Lewis and Shedler (1971) for an analysis of aspects of resource contention in multiprogrammed demand paging systems.)

The frequency and pattern of page exceptions strongly influences the performance (in fact, the feasibility) of a demand paging system. Accordingly, the study of page reference patterns and page exceptions (as a function of page size, main memory capacity, and replacement algorithm) is of interest to the system designer who must determine pertinent system resources and select system control algorithms and parameters; in particular he would like to choose a page size, replacement algorithm and main memory capacity so as (in uniprogrammed mode) to minimize the (long-run average) page exception rate. Thus it would be useful to know the stochastic structure of the reference process.

We can study several related stochastic processes in order to characterize page reference patterns:

(i) reference strings $\{R_i\}$, i.e., sequences of page references, where R_i is the page referenced by the program at time i. We can think of these as a multivariate point process (Cox and Lewis, 1972) in discrete time, the multivariate aspect being that the events (references to a page) which occur at each time instant i are of several types (different pages);

(ii) distance strings $\{D_i\}$, e.g., sequences of stack distances for least recently used (LRU) replacement, as defined in Mattson, Gecsei, Slutz and Traiger (1970), where D_i is the total number of distinct pages referenced since the last reference to R_i;

(iii) the point processes corresponding to page exceptions for various (fixed) main memory capacities c, i.e. (discrete) times i at which D_i exceeds the main memory capacity. Denote this process by $\{T_j(c); j=1,2,...\}$, where $T_j(c)$ is the time of the jth page exception in memory of capacity c. As c increases, fewer page exceptions (under LRU replacement) occur; does this thinning result in a Poisson process when c is large?

It is not necessarily simple to go (probabilistically) from one of these representations to another. It may, of course, be that one of the representations is more convenient than any of the others in a particular application. The distance string representation $\{D_i\}$ suppresses page names, which may be advantageous in that the process should be more nearly stationary than the process $\{R_i\}$.

The Lewis and Shedler (1973) paper describes analysis and modelling of page exceptions as a univariate, unmarked point process. The available data consists of the reference string $\{R_i\}$, a sequence of approximately 8.8 million references to some 517 distinct 4,096 (4K) byte pages for a particular (relatively small) program. From the reference string, the distance string was derived, and in turn the sequence of times (in number of references) between page exceptions for each of several main memory capacities. For the smallest main memory capacity, some 1807 page exceptions occur; for the largest capacity, 517 page exceptions occur. The voluminous nature of this data is characteristic of computer system data.

The initial basis for the analysis and modelling is available theory on rare events and thinning of point processes (Daley and Vere-Jones, 1972, Section 5.3) suggesting that these relatively rare events (page exceptions) should form approximately a Poisson process. An analysis of the data was undertaken to confirm or reject and extend the model. The analysis shows quickly that the Poisson model for page exceptions is grossly inadequate. A direct examination of the distance string, as well as a spectral analysis, indicates the presence of an alternation or two-state phenomenon (a consequence of so-called locality of reference), and on this basis a two-state univariate semi-Markov generated point process model (Cox, 1963, Cox and Lewis, 1966, Ch. 7) for the process of page exceptions is formulated and found to characterize the data adequately. Unfortunately fitting this model to the voluminous available data is not simple, partly because the marginal distribution of times between page exceptions is a mixture of a discrete random variable and a very skewed continuous random variable; see Figure 1. We feel that some of the new models described in the

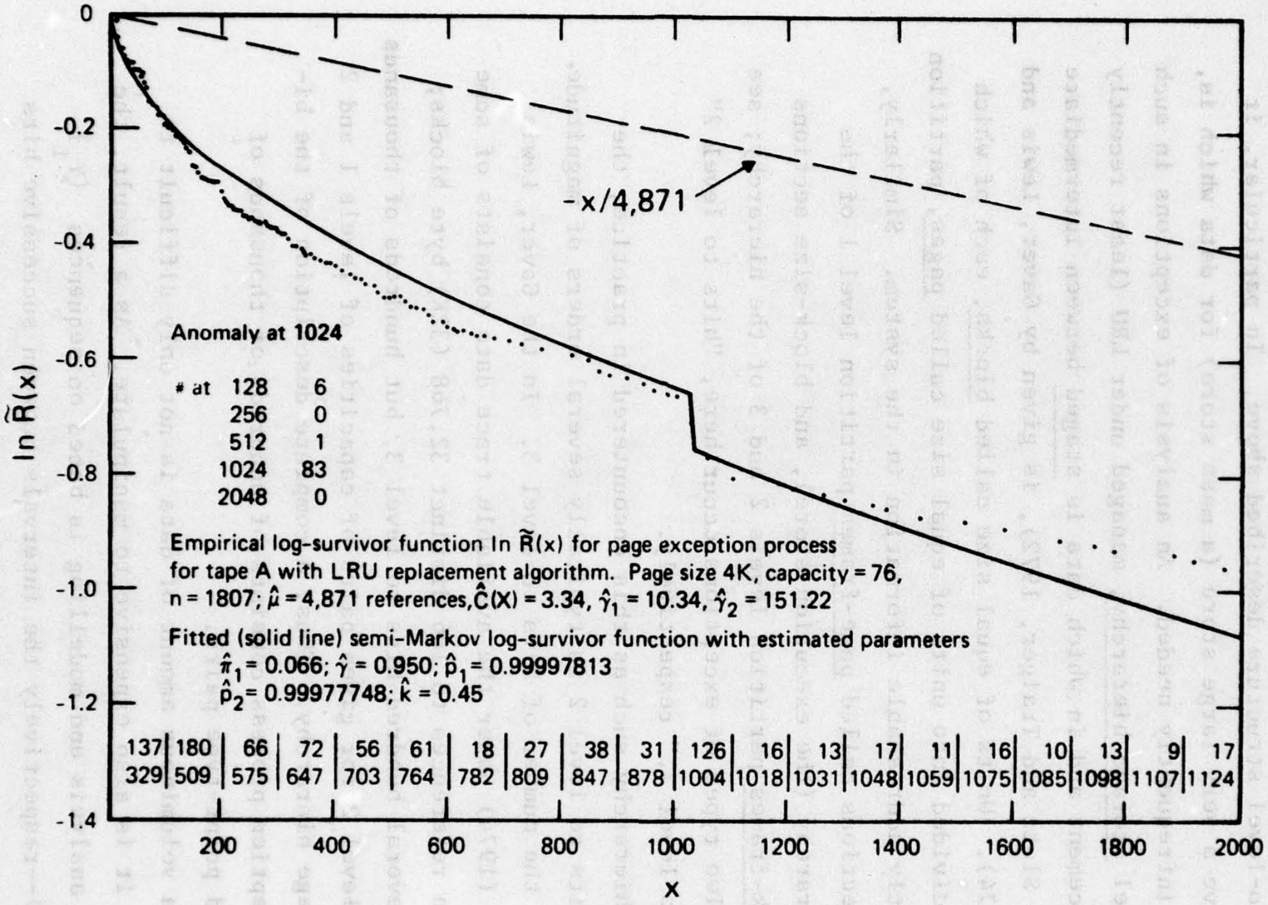

FIG. 1.

Estimated and fitted marginal distribution of intervals for page exception process; from Lewis and Shedler (1973).

section of this paper are more appropriate for this data than the univariate semi-Markov generated point process.

2.2.2. Exceptions in a three-level storage hierarchy. A demand-paged computer system may have a more general hierarchy of storage than the two-level structure described above. In particular, it may also have a very large store (a mass store) for data which is, hopefully, infrequently needed. An analysis of exceptions in such a three-level storage hierarchy, managed under LRU (least recently used) replacement and in which data is staged between intermediate levels (cf. Slutz and Traiger, 1972), is given by Gaver, Lewis and Shedler (1974). Units of equal size called blocks, each of which is further divided into units of equal size called pages, partition the explicitly addressable information in the system. Similarly, page size sections called page-frames partition level 1 of the storage hierarchy (the execution store), and block-size sections called block-frames partition levels 2 and 3 of the hierarchy; see Figure 2. Two types of exceptions occur here, "hits to level 2" and "hits to level 3," respectively.

In a hierarchy such as this encountered in practice, the number of hits to level 2 is typically several orders of magnitude larger than the number of hits to level 3. In the Gaver, Lewis and Shedler (1974) paper the available trace data consists of some 34.7 million references to 166 distinct 32,768 (32K) byte blocks; there are several hundred hits to level 3, but hundreds of thousands of hits to level 2 for given pairs of capacities of levels 1 and 2 of the storage hierarchy. Thus, a complete description of the bivariate exception process consists of hundreds of thousands of interval and point-type pairs.

Such a voluminous amount of data is not only difficult to comprehend, it is also expensive to manipulate. As a result, the statistical analysis and modelling is based on sequences $\{Y_i\}$ and $\{N(Y_i)\}$--respectively the intervals between successive hits to level 3, and the counts of hits to level 2 between successive

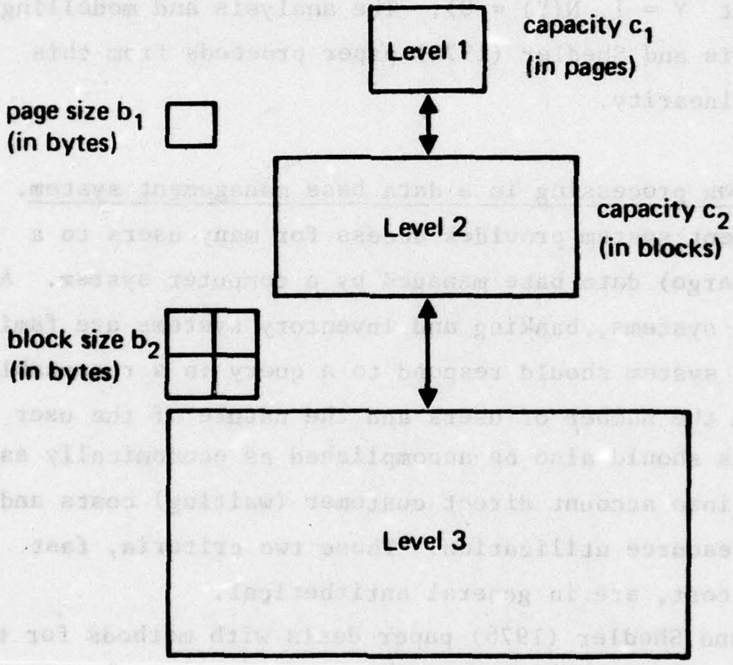

FIG. 2.

Staging of data in a three-level storage hierarchy.

hits to level 3. The starting point of the analysis is a set of scatter diagrams of points in the $[Y,N(Y)]$ plane. These scatter diagrams reveal the apparent existence of two distinct kinds of referencing behavior. The capacities of levels 1 and 2 of the storage hierarchy, in pages and blocks, respectively, are denoted by c_1 and c_2; together with b_1 and b_2, respectively, the page size and block size in bytes, these are the basic hierarchy parameters. For each pair of capacities c_1 and c_2 a striking two-line relationship is observed in the graphical display; see Figure 3. Points in the $[Y,N(Y)]$ plane appear to be of two types, and for each of the two types, points of that type cluster about a straight line (through the point $Y = 1$, $N(Y) = 0$). The analysis and modelling in the Gaver, Lewis and Shedler (1974) paper proceeds from this observed double-linearity.

2.2.3. Transaction processing in a data base management system. A data base management system provides access for many users to a (typically very large) data base managed by a computer system. Airline reservations systems, banking and inventory systems are familiar examples. Such a system should respond to a query in a reasonably short time, given the number of users and the nature of the user environment. This should also be accomplished as economically as possible, taking into account direct customer (waiting) costs and computer system resource utilization. These two criteria, fast response and low cost, are in general antithetical.

The Lewis and Shedler (1976) paper deals with methods for the examination of nonstationary univariate point processes which can be applied to obtain a graphical and mathematical description of the behavior of a running data base management system. Such a description provides a useful starting point for studies aimed at workload characterization, a central problem in performance evaluation of data base management systems. Stochastic models of the kind obtained by Lewis and Shedler (1976) have application to the detailing of proposed (e.g., queueing network) system models and to the validation of such system models.

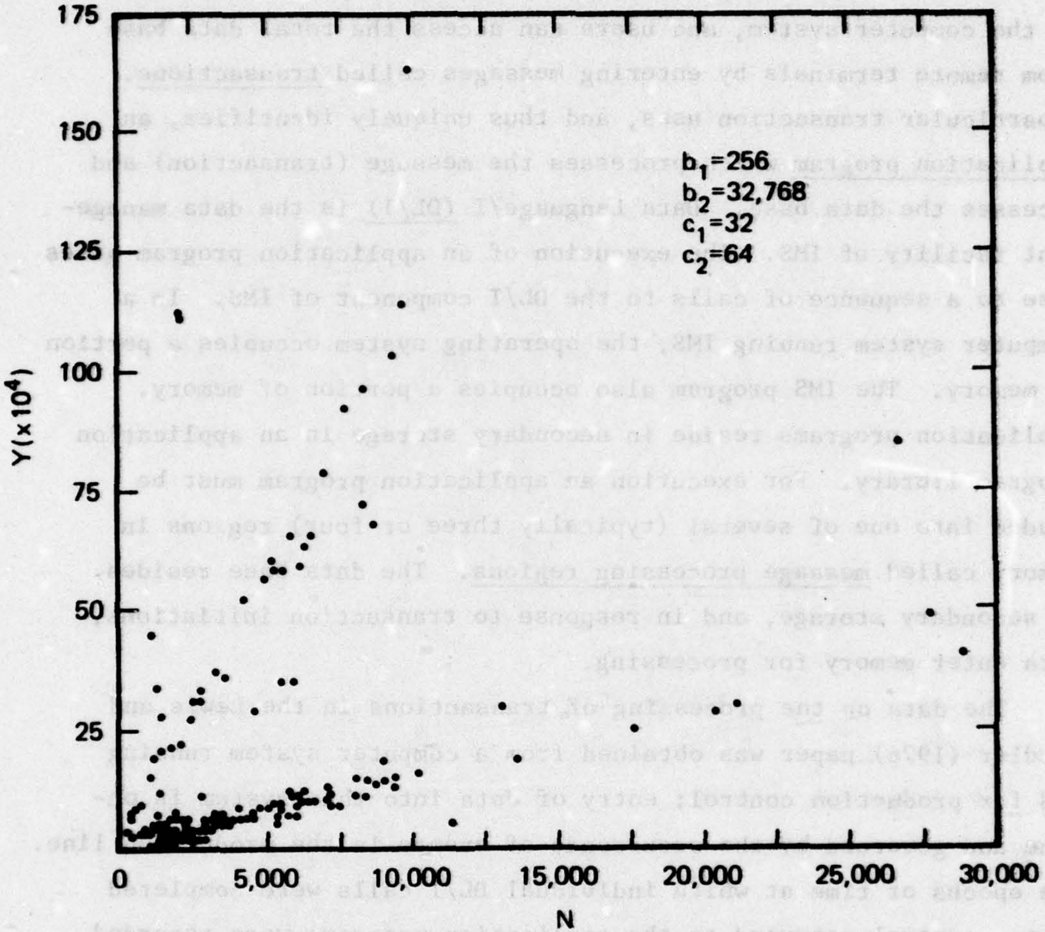

FIG. 3.

Scatter diagram for bivariate exception process in a three-level storage hierarchy; from Gaver, Lewis and Shelder (1974).

The analysis by Lewis and Shedler (1976) is of data obtained from an IMS (Information Management System) data base management system. IMS (IBM Corp., 1973) is a processing program for the implementation of large data bases accessed in common by several applications. The IMS program executes under the operating system of the computer system, and users can access the total data base from remote terminals by entering messages called transactions. A particular transaction uses, and thus uniquely identifies, an application program which processes the message (transaction) and accesses the data base. Data Language/I (DL/I) is the data management facility of IMS. The execution of an application program gives rise to a sequence of calls to the DL/I component of IMS. In a computer system running IMS, the operating system occupies a portion of memory. The IMS program also occupies a portion of memory. Application programs reside in secondary storage in an application program library. For execution an application program must be loaded into one of several (typically three or four) regions in memory called message processing regions. The data base resides in secondary storage, and in response to transaction initiations, data enter memory for processing.

The data on the processing of transactions in the Lewis and Shedler (1976) paper was obtained from a computer system running IMS for production control; entry of data into this system is on-line and governed by the occurrence of events in the production line. The epochs of time at which individual DL/I calls were completed (i.e., control returned to the application program) were recorded, along with information sufficient to identify the initiation times of individual transactions.

In analyzing the transaction initiation data, there are a number of prior assumptions that can be made about the data to serve as a starting point. The purpose of the data analysis is to confirm these assumptions or to point to suitable modifications. Since the data was taken over six whole days (typically some 25,000 transactions per day), a time-of-day effect would be expected as activity builds up through the working day and then declines during the evening. Thus any kind of initial analysis based on an assumption of stationarity is inappropriate. The usual null model is a non-homogeneous Poisson process; this could be reasonable since the transaction initiation process is a superposition of inputs from a number of sources (users). Because each user's activity is likely to consist of a (random) number of transactions after initial sign-on, some clustering in the data might be expected. An appropriate model here is the non-homogeneous branching Poisson process or Poisson cluster process (Lewis, 1967). In this process an initial primary (main) event generates a finite sequence of secondary (subsidiary) events; the complete process is then the superposition of the primary and secondary events, where a non-homogeneous Poisson process generates the main events. If there are enough primary events (high-activity) so that the number of active secondary processes is large, the process is hard to distinguish from a non-homogeneous Poisson process.

Starting from these assumptions, the analysis of the data proceeds as follows. A very rough, model-free procedure gives an estimate of the rate function for the transaction initiation process over the whole day. On the basis of this trend analysis, relatively homogeneous high- and low-activity periods during the day are selected, and an attempt is made to verify the non-homogeneous Poisson process model or the cluster process model. Based on this local analysis and modelling of the transaction initiation process, more formal model-dependent procedures are applied to the transaction rate function for the several days. The Poisson assumption is found to be reasonably valid for high-activity periods; clustering becomes more evident in low-activity periods.

3. STATISTICAL ANALYSIS OF POINT PROCESSES IN COMPUTER SYSTEMS

3.1. The Nature of Computer System Data

There are five important characteristics of data obtained during the measurement phase of computer system performance studies.

(i) The amounts of data available are staggering; often several sequences or series of events are observed for times producing millions of observations.

(ii) The times-between-events are often overdispersed relative to an exponential distribution, and in many cases contain discrete components.

(iii) Stationarity is often not a reasonable assumption; frequently when stationarity is reasonable it is because there is random switching back and forth between several possibly stochastic modes.

(iv) There are many situations in which there are fairly gross inhomogeneities in the data; these can usually be tied to external variables such as the number of users of the system, or to the time of day.

(v) Sometimes the data suggests that there may simply be no stochastic regularity involved. Of course, it could be that, in line with (iii), not enough time is involved to show emergent patterns.

As a consequence of (v) there is a procedural difficulty. Consider starting a rough exploratory analysis of a series of events by smoothing the data to obtain a graph of the event rate over time. Suppose we observe that over the first 100,000 events the rate is fairly constant and that beyond this it changes, a phenomenon which can be seen by eye and validated by simple statistical methods. Should we analyze the data in two parts, and if so, how do we characterize the process in toto? Should we take more data and hope to distinguish "all" the possible stochastic modes?

On the other end of the scale, when we examine the apparently stationary segments of the data, questions of more microscopic stationarity sometimes arise. This is reminiscent of the self-similarity concept for physical data put forward by Mandelbrot (1967).

An unfortunate consequence of all of the above is that we can seldom sample the data over time to achieve some compression. Even in the best of circumstances, intelligent use of sampling techniques requires well-formulated questions; the tendency in computer system measurement often is to feel that if an enormous amount of data is collected, it will be possible to answer any question that may arise!

3.2. Recent Developments

Lewis (1972) summarizes some of the recent developments in the statistical analysis of point processes; see also Cox (1972) and Brown (1972). There is also a book on point processes by Snyder (1975) and a sequence of papers by Brillinger (1975a, 1975b). Brillinger bases his work heavily on spectral methods for stationary processes and his work has many points of contact with that of Cox and Lewis. But Snyder (1975) does not reference any of Brillinger's papers, and does not reference Cox and Lewis (1966). What then is the point of contact between these lines of development in the analysis of series of events?

The work of Cox and Lewis and that of Brillinger are fairly complementary. The former is highly data analytic in the sense that there is concern with validation of assumptions and models and analysis of trends; the latter is concerned mainly with spectral methods for stationary (univariate and multivariate) point processes based on models such as self-exciting point processes (Hawkes, 1972, Hawkes and Oakes, 1974) which lend themselves easily to spectral methods. Snyder (1975) bases his statistical methods almost entirely on likelihood analysis, a strong assumption that the stochastic mechanism generating the series of events is known and that it is possible to write down the "sample-function density." We have doubts that this approach will be useful in analyzing data from computer systems, in particular since for this type of data there do not appear to be any compelling physical models other than, in some cases, Poisson and Poisson cluster processes.

Consider now some specific areas of development.

3.2.1. Trend analysis and detrending of point processes. In many fields of application, and particularly in computer system data, it has become increasingly apparent that stationary point processes are at best a convenient mathematical fiction. Most data exhibit fairly subtle trends and methods for testing for these trends are known (Cox and Lewis, 1966, Ch. 3); other data, however, exhibit gross trends, e.g., time-of-day effects in the series of arrivals at a queue, and techniques for the analysis and characterization of such data are only now beginning to be developed.

The situation is analogous to that in ordinary regression analysis and time-series analysis where we might want, for example, to estimate parameters in an assumed (linear) function for the mean, test the model for the mean function and then examine the model which is assumed for the residuals. The latter could include examining the residuals to test for independence, estimating the spectrum of the residuals and testing the assumed normality of the residuals. Techniques for these problems in the linear normal model are known (see e.g., Hannan, 1970).

By comparison, in point processes we might want to:

(i) estimate the rate function $\lambda(t)$, using either specific functional models or smoothing techniques;

(ii) test specific functional models for $\lambda(t)$;

(iii) detrend the point process, examine the 'residual' process and test the usual hypothesis that the events are generated by a homogeneous Poisson process.

When dealing with non-homogeneous Poisson processes, the most appropriate detrending technique (Lewis, 1970, 1972) seems to be to transform the time-scale so that the ith event occurring at time t_i now occurs at time

$$\tau_i = \int_0^{t_i} \hat{\lambda}(u)du ,$$

where $\hat{\lambda}(t)$ is some estimate of the rate function $\lambda(t)$. Note that if $\lambda(t)$ is known, the $\{\tau_i\}$ process is a Poisson process of rate one. When $\lambda(t)$ is estimated from the data, the distributional problems associated with determining properties of the $\{\tau_i\}$ process are difficult.

Results for estimating parametric rate functions are given by Cox (1972) and Lewis (1972). These methods are developed by Lewis and Shedler (1976) and applied to the statistical analysis of transaction processing in the data base systems. It is probably best to deal with fairly regular point processes by using log transforms of the intervals between events and then using ordinary time-series methods (Cox and Lewis, 1966, Ch. 3). It is still difficult to deal with non-homogeneous processes which are overdispersed relative to a Poisson process, e.g., a non-homogeneous Poisson cluster process; some work has been done by Lewis and Robinson (1974).

3.2.2. Spectral analysis of point processes. By spectral analysis of a point process (Barlett, 1963) we mean the second-order spectrum of the counting function $N(t)$ of the point process (the count spectrum). Brillinger (1972) has put this spectral analysis on a firm footing in the context of a general spectral theory for stationary interval functions such as $N(t)$. He has also proposed the use of higher-order spectra.

We can think of the spectral analysis of a point process as an ordinary second-order spectral analysis of a function $dN(t)$ which is a series of delta functions occurring at random times $\{T_i\}$, the times-to-events; see Lewis (1970) for a heuristic interpretation. Note that the second-order count spectrum completely specifies a renewal process. This spectrum, $g_+(\omega)$, is, in fact, essentially the Fourier transform of the renewal density or the intensity function. Note that this spectral analysis is not a second-order spectral analysis of the intervals between events $X_i = T_i - T_{i-1}$. The latter spectrum, denoted by $f_+(\omega)$, is useful for differentiating between renewal processes (for which it is flat) and non-renewal point processes. The spectrum of intervals may, in fact, be preferable to higher-order spectra of counts (Brillinger, 1972) in that it should exhibit fewer sampling fluctuations; in general our feeling is that a spectral analysis of the counts and the intervals should be tried before going to higher order spectra. The estimated spectrum of intervals $\tilde{f}_+(\omega)$ for the page exception process of Lewis and Shedler (1973) is shown in Figure 4. The underlying spectrum $f_+(\omega)$ is clearly not flat (i.e. equal to 2 for all ω) so that the process is neither a Poisson process nor a renewal process. The spectrum of a mixed moving average-autoregressive process (ARMA(1,1)), where the orders of the moving average and the autoregression are both one, fits the estimated spectrum well. We return to this in the next section; the two-state semi-Markov generated point process model used by Lewis and Shedler (1973) and several of the models defined in Section 4 have this spectrum. The fitted interval spectrum is shown in the figure. The estimated spectrum of counts, along with the fitted spectrum of counts of the two-state semi-Markov generated point process for the page exception process data of the Lewis and Shedler (1973) paper are shown in Figure 5.

One drawback to the spectral analysis of point processes is the large amount of time required for computation of spectral estimates. Only recently have French and Holden (1971), in an important paper, found a way to use the fast Fourier transform (FFT) in this context. There are some problems with this technique, e.g., it is not bias-free but it appears to be of great value.

We shall return to Brillinger's higher-order count spectra in the discussion of new models. In most cases involving computer system data there is a problem in applying spectral techniques because of lack of stationarity. Used with care, however, spectral techniques can indicate a switching of levels or some kind of quasi-cyclic behavior in a system (see e.g., Lewis and Shedler, 1973).

3.2.3. Multivariate point processes. In almost all applications in computer systems we are interested in interactions between

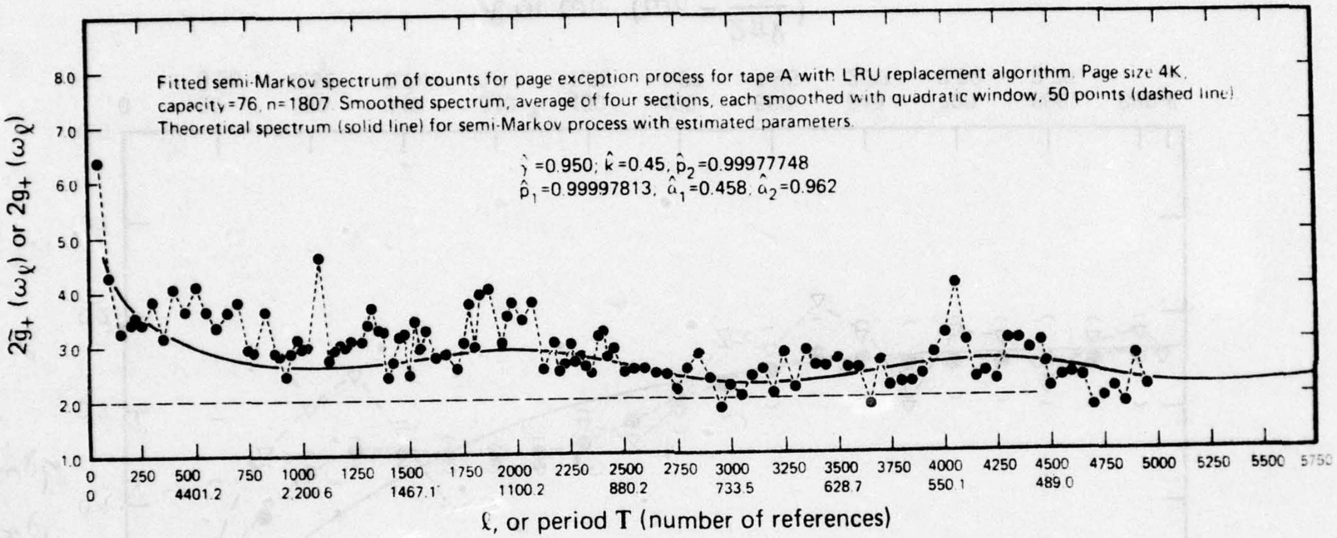

FIG. 4.

Estimated and fitted spectrum of intervals for page exception process; from Lewis and Shedler (1973).

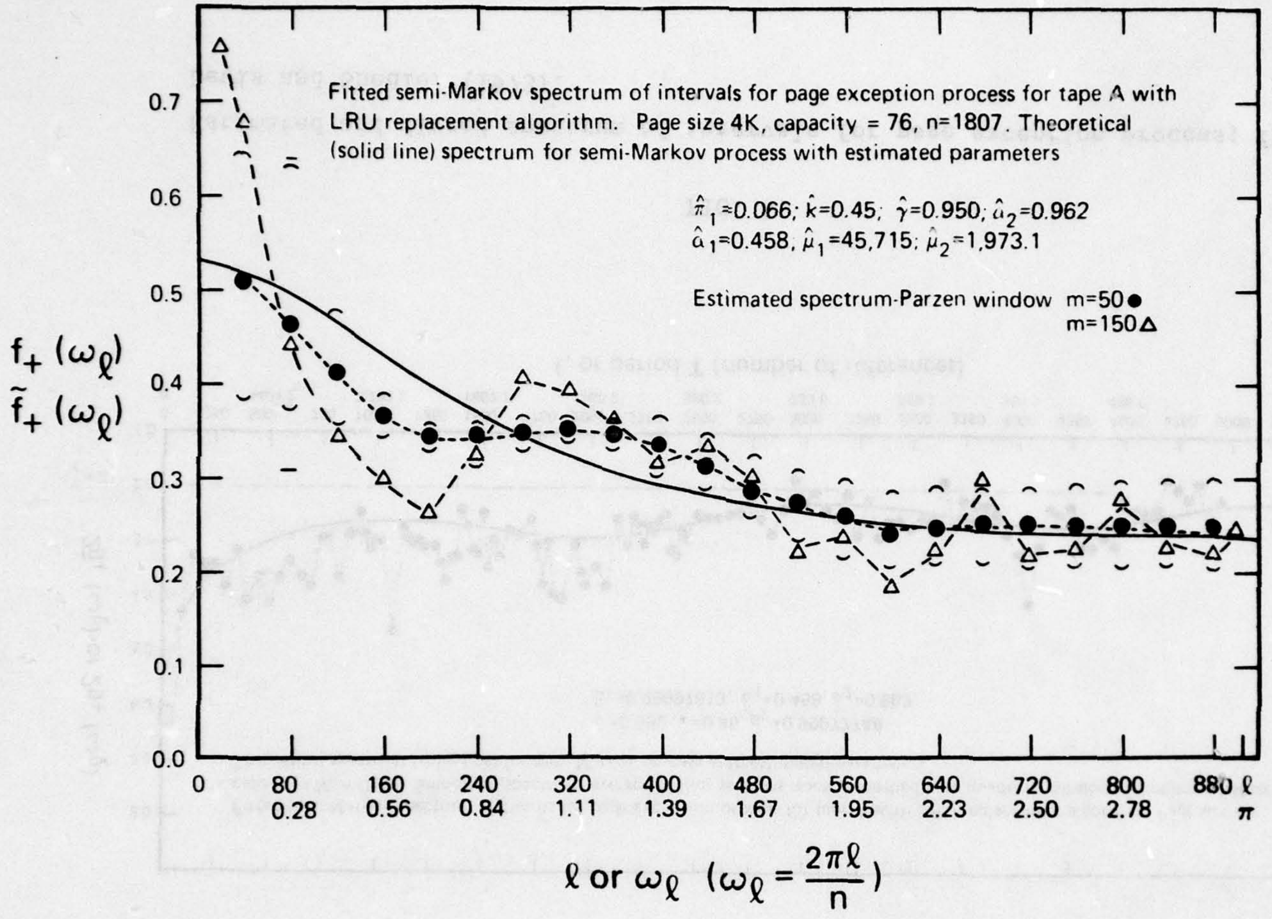

FIG. 5.

Estimated and fitted spectrum of counts for page exception process; from Lewis and Shedler (1973).

stochastic processes occurring at different places in space and time. We have given some examples above; additional examples are the following:

(i) arrivals of requests to the several spindles of a disk storage device;

(ii) times of occurrence of references to different programs in multiprogrammed computer system;

(iii) times of job start and termination by region in a multiprogrammed computer system (Hunter and Shedler, 1977).

These examples are by no means exhaustive but illustrate important applications. We shall discuss multivariate point processes (Cox and Lewis, 1972) which we think of as point processes in which qualitative marks associated with each event partition the sets of events. Note that Brillinger's work (Brillinger, 1972) encompasses more general situations. Cox and Lewis (1966, Ch. 8, 1972), Perkel, Gerstein and Moore (1966) and Brillinger (1972, 1975a, 1975b) discuss the analysis of multivariate point processes, and we will make only general comments here.

a) Dependencies between two stationary processes are usually handled via spectral methods, i.e., second-order cross-spectra which when normalized give quantities called coherences. This is Brillinger's approach. It is not at all clear, however, how useful second-order spectra are for point processes (univariate and multivariate), which are a long way from normal processes in which the second-order spectra completely specify the dependence.

b) There is a problem of specifying what kind of dependency structures occur in multivariate point processes. We can, for instance, generate many bivariate Poisson processes, i.e., bivariate point processes in which the individual (marginal) processes are Poisson. Cox and Lewis (1972) give a start at examining these structures; also see Lawrance and Lewis (1975) and Oakes (1976).

c) The above leads to questions of suitable models for multivariate point processes; Cox and Lewis (1972) address this in a tentative way, but there is much more to be done. Mutually exciting point processes (Hawkes, 1972) are being used, but are not well understood and are, we believe, very untractable. Some recent work (Lawrance and Lewis, 1977, Jacobs and Lewis, 1977a) discussed below, when extended to multivariate processes, gives promise of simpler models.

d) Spectral methods are only one of many tools for examining dependencies in point processes and we have already expressed our reservations about their use for this problem. There is a paper by Cox (1972) which is important and offers techniques based on likelihoods which need further exploration. They might, in particular, be useful for analyzing dependencies in non-stationary point processes such as considered by Lewis and Shedler (1976). Note that spectral methods are not applicable to non-stationary point processes.

e) Finally, the value of graphical data analytic methods should be appreciated and their use explored before shotgun methods (e.g., spectral analyses) are used. This is particularly true in applications to computer system data since it is not completely evident that there is always stable stochastic structure in the processes encountered.

An example of a very fruitful, simple graphical analysis of a bivariate point process (discussed above) appears in Gaver, Lewis, and Shedler (1974); the plot in Figure 3 is striking in that it consists (within statistical fluctuations) of two separate straight lines. The main thrust of the paper is to model this phenomenon. This illustrates the way data analysis should be used: to suggest models or modify postulated models.

4. MODELS AND MODELLING OF POINT PROCESSES IN COMPUTER SYSTEMS

In the previous section we discussed characteristics of point process data observed in computer systems; here we discuss some aspects of modelling point processes of this kind. One reason for discussing modelling in this context is that the peculiarities of the data have, in a fairly insistent way, led us to develop the new models described below. Why do we need new models, in particular, to describe the internal complexities of a computer system? The physics and data analysis of the computer reliability problem led to an important model (the Poisson cluster process) which has application in many other contexts. There is, however, usually no such physical imperative in the internal computer processes, and the data analysis typically reveals enormous complexity which is difficult to match to characteristics of the usual point process models (e.g., cluster-processes, doubly stochastic Poisson processes, etc.). Besides non-stationarity, which we ignore here, complexity of the modelling is apparent from an analysis of data on the marginal distribution of times-between-events. In the analysis of page exceptions given by Lewis and Shedler (1973), the marginal distribution is found to be highly skewed and to have a discrete component (Figure 1). None of the common point process models can describe the marginal distribution of such data, let alone its dependency structure.

Models for computer system performance evaluation have the following requirements:

(i) First, there is a need for descriptive and structurally simple point process models analogous to the linear processes used in the usual time-series analyses (e.g., Box-Jenkins techniques). These should be easy to fit to the data, and simple to generate on a computer, since the models are often used in simulation studies of computer system performance.

(ii) Second, there is a need for models in which the marginal distribution of times-between-events is specifiable in a manner which is as independent of the specification of the dependency structure of the model as is possible.

With regard to the second point, the computer system data studies described in this paper have made us aware of the extent to which the usual point process models are primarily concerned with the dependency structure of the model. (The analog in ordinary time series analysis is that in defining linear models we usually assume that the random variables are normally distributed, or ignore this aspect of the problem altogether.) The distribution of times-between-events, however, is one of the most easily observed aspects of a point-process, and can be just as informative as, say, the spectrum of counts. The estimated marginal distribution of the page exception process given in Figure 1 has a discrete component at $x = 1024$; this artifact of the data can be related in an informative way to the paging process.

We describe now some recently developed stochastic sequences which are useful as models for point processes. We intend no implication that the constructions are unique. The sequences do have properties, however, which make them very useful in modelling point processes in computer systems. In particular, the marginal distribution of the variables is an integral part of the specification of the stochastic sequence.

4.1. Interval Models.

Univariate point processes can be described equally well through the structure of the intervals between events $\{X_i\}$ or the counting process $\{N(t)\}$, where $N(t)$ gives the number of events in $(0,t]$. We discuss the modelling of the intervals $\{X_i\}$ first.

4.1.1. The first-order autoregressive exponential model (EAR1). In a Poisson process the intervals $\{X_i\}$ are independent and identically distributed (i.i.d.) with exponential (λ) distribution

$$F_X(x) = 1 - e^{-\lambda x}\ , \quad \lambda > 0;\ x \geq 0\ . \qquad (4.1)$$

Several attempts have been made to generalize the Poisson process by making the X_i dependent, but with exponential or conditionally exponential marginal distributions (Cox, 1955). The simplest and only really successful attempt in the sense of broad applicability (Gaver and Lewis, 1977), gives a process called the EAR1 model, derived from the following consideration.

A first-order autoregressive stochastic sequence is defined by the stochastic difference equation

$$X_i = \rho X_{i-1} + \varepsilon_i \,, \qquad i = 0, \pm 1, \pm 2, \ldots ; \; |\rho| < 1 \,, \tag{4.2}$$

where the ε_i are assumed to be an i.i.d. stationary random sequence. If the ε_i are normally distributed, so are the X_i. What must the distribution of the ε_i be in order for the X_i sequence to be stationary with an exponential (λ) distribution? The answer is surprisngly easy (Gaver and Lewis, 1977).

Let $0 \leq \rho < 1$, and $\{E_i\}$ be an i.i.d. exponential (λ) sequence. Now let ε_i be equal to zero with probability ρ and equal to E_i with probability $1-\rho$. Then we have

$$X_i = \begin{cases} \rho X_{i-1} & \text{probability } \rho \\ \rho X_{i-1} + E_i & \text{probability } (1-\rho) \end{cases} \tag{4.3}$$

$$= \rho X_{i-1} + V_i E_i \,, \tag{4.4}$$

where $\{V_i\}$ is an i.i.d. binary sequence with $V_i = 1$ with probability $(1-\rho)$. Moreover if we let $X_0 = E_0$, and define X_i as in (4.3), the resulting sequence is stationary for $i = 0, 1, \ldots$.

The point process with the interval structure (4.3) is called the EAR1 point process. It is a tractable model, and most of its important properties are given in Gaver and Lewis (1977). In particular we have that $\rho(k) = \rho^k$. This model is in a sense degenerate because it contains runs of X_i in which values are exactly ρ times the previous value; it could, however, be a reasonable model for point processes observed in computer systems (e.g., inter-arrival times of requests to a storage subsystem) in which the intervals have exponential marginal distributions but are dependent. Note that as defined the model can only provide sequences $\{X_i\}$ with positive serial correlations. We can, however, define the process to include negative correlations.

Simple generalizations of this Markovian exponential process are the following.

4.1.2. The moving average exponential model (EMAk). We define another stationary sequence $\{X_i\}$, using the $\{E_i\}$ sequence above, according to

$$X_0 = E_0 \qquad (4.5)$$

$$X_i = \beta E_i + U_i E_{i-1}, \qquad i = 1,\dots ; \quad 0 \le \beta \le 1 , \qquad (4.6)$$

where $\{U_i\}$ is an i.i.d. binary sequence in which $U_i = 1$ with probability $(1-\beta)$. This is a first order exponential moving average process (EMA1) (Lawrance and Lewis, 1977) which is one-dependent; in particular

$$\rho(1) = \beta(1-\beta) \qquad (4.7)$$

$$\rho(k) = 0 , \qquad k = 2,3,\dots . \qquad (4.8)$$

Properties of the EMA1 process are given by Lawrance and Lewis (1977).

It is easy to see that we can make E_{i-1} in (4.6) a random linear combination of E_{i-1} and E_{i-2} to get an EMA2 process, and can continue the process back k steps to obtain an EMAk process. In addition, by making E_{i-k} autoregressive over the previous E_i, we obtain a mixed kth order moving-average, first order autoregressive process which we denote by EARMA(1,k).

4.1.3. The EARMA(1,1) model. Consider explicitly the case $k = 1$. The first order moving-average and first order autoregressive process EARMA(1,1) is given by

$$X_i = \beta E_i + U_i A_{i-1} \qquad (4.9)$$

with

$$A_{i-1} = \rho A_{i-2} + V_i E_{i-1} \qquad (4.10)$$

for $i = 1, 2, 3, \ldots$ and $A_{-1} = E_{-1}$. This sequence of random variables is not Markovian.

The second-order correlation structure of the process is given by

$$\rho(k) = \rho^{k-1} c(\beta,\rho) \ , \tag{4.11}$$

where

$$c(\beta,\rho) = \beta(1-\beta) + \rho(1-\beta)(1-2\beta) \ . \tag{4.12}$$

The point process whose intervals have the EARMA(1,1) structure is discussed in detail in Jacobs and Lewis (1977a). In particular, for $\beta = 1$ it is a Poisson process. The process is very simple to generate on a computer and is very useful for modelling dependent sequences in queuing systems. It is possible to give an extenstion to processes in which the X_i are Gamma distributed, but not much beyond this. In fact a necessary condition to ensure that we can find an ε_i in the fundamental relationship (4.2) to give a specified distribution $F(x)$ for X_i is that $F(x)$ be infinitely divisible.

We discuss now a possibly broader but more complex model for point processes having a specified interval *distribution*.

4.1.4. <u>The semi-Markov generated point process with fixed marginal distribution</u>. The question arises as to whether there are interval processes $\{X_i\}$ with exponential marginal distributions and ARMA(1,1) second-order correlation structure and which cover a broader range of correlation than the EARMA(1,1) process (though perhaps at a cost of more complicated structure).

We discuss briefly one such process. It is a special case of the semi-Markov generated point process introduced by Cox (1963) and extended by Haskell and Lewis (1977). We first describe the two-state semi-Markov generated model. In this model there are two types of intervals with distributions $F_1(x)$ and $F_2(x)$, sampled in accordance with a two-state Markov chain for which the one-step transition matrix

$$\underset{\sim}{P} = \begin{pmatrix} \alpha_1 & 1-\alpha_1 \\ 1-\alpha_2 & \alpha_2 \end{pmatrix} \tag{4.13}$$

and

$$\underset{\sim}{\Pi} = \underset{\sim}{\Pi}\underset{\sim}{P} = \left(\frac{1-\alpha_2}{2-\alpha_1-\alpha_2} , \frac{1-\alpha_1}{2-\alpha_1-\alpha_2} \right). \tag{4.14}$$

When we form the point process we assume that no information is available about the type of interval, i.e., that in the actual bivariate point process of transitions we suppress knowledge of the type of transition. Then the distribution of an interval between transitions (events) X_i in the stationary point process is

$$F_X(x) = \pi_1 F_1(x) + \pi_2 F_2(x) \tag{4.15}$$

and the correlation between X_i and X_{i+k} is

$$\rho(k) = M\beta^k , \qquad k = 1,2,\ldots , \tag{4.16}$$

where M is a positive constant and $\beta = \alpha_1 + \alpha_2 - 1 = \alpha_1(1-\alpha_2)$. Thus the correlation structure is that of an ARMA(1,1) process. For a derivation of this result see Cox and Lewis (1966), Ch. 7, 194-196. Lewis and Shedler (1973) use this process to model the page exception process. The problem is to deal with the mixture distribution (4.15) for the marginal distribution of intervals; this seems to limit the utility of the model.

To obtain an exponential marginal distribution, consider the following device (Jacobs and Lewis, 1977a). Fix x_0, where $0 < x_0 < \infty$, and let

$$F_1(x) = \begin{cases} \dfrac{\int_0^x \lambda e^{-\lambda u} du}{1 - e^{-\lambda x_0}} & 0 \le x \le x_0 , \\ 1 & x > x_0 ; \end{cases}$$

$$F_2(x) = \begin{cases} 0 & x \le x_0 , \\ \dfrac{\int_{x_0}^x \lambda e^{-\lambda u} du}{e^{-\lambda x_0}} & x > x_0 ; \end{cases} \qquad (4.17)$$

then $F_X(x)$, the marginal distribution of an interval, is exponential (λ) if we set $\pi_1 = 1 - \exp(-\lambda x_0)$. There is one degree of freedom left in the matrix $\underline{P}$; in addition to λ, we have free parameters π_1 (or x_0) and α_1. What then is the range of β, and can it be negative?

Straightforward manipulation shows that

$$\beta = \frac{\pi_1 - \alpha_1}{\pi_1 - 1} , \qquad (4.18)$$

which lies in absolute value between zero and one but can be negative; therefore the serial correlations can be negative. Thus the model appears to be broader than the EARMA(1,1) model. The question of comparing the two models when β is positive has not yet been explored; it requires higher order interval correlations, as discussed by Brillinger (1972).

By letting $F_1(x)$ and $F_2(x)$ be a partitioning as in (4.17) of any specified distribution $F(x)$ we obtain a point process whose marginal interval distribution is the specified distribution $F(x)$ (discrete, continuous or mixed), and which has ARMA(1,1) type second-order interval spectrum and known count spectrum. We note that there is a choice of x_0 which gives a geometrically decaying $\rho(k)$, but unlike the EAR1 process, the resulting process is not Markovian.

By performing the same type of truncation on an n-state semi-Markov generated point process (Haskell and Lewis, 1977) it is possible to obtain a point process with specified marginal distribution, almost any ARMA-type second-order interval correlation structure (i.e., spectra which are ratios of polynomials in cos ω) and known count spectrum. In fact this seems to be the only point process model for which all these characteristics are known and easily computable. Properties of this model have not yet been fully explored. The one disadvantage of this model viz-a-viz the exponential models described in the previous section is that, since the model is not a probabilistic linear combination of random variables, it is not easy to relate to intuitive considerations when used in computer system models. We return to this aspect of the modelling below when we discuss multivariate point processes.

4.2 Models for Counts

It is not always possible to observe the exact times of events in a point process and in fact,with respect to computer system data, such data gathering can be very expensive. What is more usual is to observe the counts of events in successive intervals of a fixed length Δ. We denote the differential counts of events in successive intervals by N_i, $i = 0,1,\ldots$. To model the $\{N_i\}$ sequence we need, in general, models for dependent sequences of positive valued, discrete random variables. Of course if we observe a Poisson process the $\{N_i\}$ are independent and Poisson distributed. Otherwise, we know of no model, defined in terms of exact occurrences of events, for which the characteristics of the N_i process are simple or known.

The modified semi-Markov generated sequence of Section 4.1.4 yields a simple model for counts by letting $F(x)$ be a discrete distribution. It would be interesting to see how closely we can approximate the differential count process of a Poisson cluster process this way.

An even simpler model for counts follows. Its main drawback is that, as defined, only positive correlations are representable.

4.2.1. The discrete mixed autoregressive-moving average DARMA(1,N+1) process. Although analogous in definition to the EARMA process, this process is very different in structure and much broader. Let the sequences $\{U_i\}$ and $\{V_i\}$ be as above, and $\{E_i\}$ be an i.i.d. sequence with any distribution $\Pi(x)$. Then the DAR1 process defined by

$$N_i = V_i N_{i-1} + (1-V_i)E_i$$

is a first-order Markov process in which the N_i have distribution $\Pi(x)$. Since successive values of the N_i can be identical, the model is useful for discrete valued processes such as the differential count process $\{N_i\}$. The process is a Markov chain with transition probabilities

$$P\{N_{i+1} = \ell | N_i = k\} = P(k,\ell) = \begin{cases} (1-\rho)\ \pi(\ell) & \text{for } k \neq \ell, \\ \rho + (1-\rho)\ \pi(\ell) & \text{for } k = \ell. \end{cases}$$

Observe a difference from the usual Markov chain modelling. The marginal distribution $\Pi(x)$ of the N_i is specified first and then the dependency structure is specified by the single parameter ρ. The model has the same drawback as the EAR1 model; the correlations are all positive, although this is not an enormous drawback when analyzing sequences of positive valued random variables.

It is possible to generalize the model to give a mixed moving-average autoregressive dependency structure. This generalization is the DARMA(1,N+1) model in Jacobs and Lewis (1977b, 1977c) defined as follows:

Let $\{Y_i\}$ be a sequence of independent real valued random variables having a common distribution π. Let $\{U_i\}$ and $\{V_i\}$

be independent sequences of $\{0,1\}$ random variables such that

$$P\{U_i = 1\} = \beta \quad \text{and} \quad P\{V_i = 1\} = \rho\,,$$

where β and ρ are fixed constants with $0 \leq \beta \leq 1$ and $0 \leq \rho < 1$. Finally, let $\{S_n\}$ be a sequence of independent random variables taking values in $\{0,1,\ldots,N\}$ with distribution F, where N is a fixed non-negative integer. Let

$$X_i = U_i Y_{i-S_i} + (1-U_i)A_{i-N-1}\,, \qquad i = 1,2,\ \ldots\ ,$$

where

$$A_i = V_i A_{i-1} + (1-V_i)Y_i\,, \qquad i = -N,-N+1,\ldots\,.$$

Perhaps the most interesting characteristic of the model is that if we transform the variables N_i, the resulting process has the <u>same</u> dependency structure as the $\{N_i\}$ process. This is because the model is a mixture of random index model and each N_i is a randomly chosen member of the $\{E_i\}$ sequence. This model therefore gives the ultimate in independence of the marginal distribution and the dependency structure. In this and other ways it is very much the analog of the normal linear processes.

Although we have introduced the DARMA(1,N+1) sequence as a model for the differential count process, it has also been used in Shedler (1977) to model sequences of event marks in multivariate point processes. In this context, event types generally provide qualitative information about the multiprogrammed processing of jobs (e.g. job start, job termination, jobstream identity) whereas event marks provide quantitative workload information.

4.3. Multivariate Processes and Systems Modelling.

The use of multivariate point process models in computer system evaluation studies is quite recent. Hunter and Shedler (1977) have defined particular marked multivariate point process models and used them for the prediction of response times in multiprogrammed systems. To illustrate another approach, we discuss use of the exponential processes EARMA(1,k) to model a single-server first-come-first-served queue in which the service times and interarrival times have exponential marginal distributions. We choose this queueing structure for simplicity of exposition; it illustrates the power of the random-linear structure of the EARMA(1,k) model in modelling queues with dependence. Moreover, it is possible to use the technique to incorporate realistic workload characteristics into networks of queues used as models for the structure of computer systems. We can also use the resulting bivariate process of service and interarrival times as a model for a bivariate point process or a highly correlated univariate process in which there are quantitative marks associated with each event.

Let S_i denote the service time for the ith arrival at the queue, and X_i denote the time between arrival of the ith and (i-1)st customer. If these are i.i.d. exponential random variables with parameters λ and α respectively, we have an M/M/1 queue.

Now for $i = 0, \pm 1, \ldots$, let $\{E_i\}$ be exponential (λ) and independent, and $\{\mathcal{E}_i\}$ be exponential (α) and independent. In addition the $\{E_i\}$ and $\{\mathcal{E}_i\}$ sequences are mutually independent. We want a queue with autocorrelated and cross-correlated service and arrival times such that it gives the M/M/1 queue as a special case, and proceed as follows.

Let $\{S_i\}$ be an EARMA(q,k) process over $(E_i, \frac{\alpha}{\lambda}\mathcal{E}_i, \frac{\alpha}{\lambda}\mathcal{E}_{i-1}, \ldots)$ where $q = 0$ or 1, and $k = 0,1,2,\ldots$. Then if $X_i = \mathcal{E}_i$, $i = 0, +1, \pm 2, \ldots$, we have that $\{S_i\}$ is EARMA(q,k) and also cross-correlated with $X_i = \mathcal{E}_i$; although $\{X_i\}$ is still a Poisson process, $\{S_i, X_i\}$ is a bivariate sequence of random variables with exponential marginal distributions.

More general schemes are possible, but the above scheme has the following simple interpretation. We have positive correlation between S_i and, most particularly, the previous k interarrival times. If the $\mathcal{E}_j$, and consequently the X_j, $j = i, i-1, \ldots, i-k-1$, are short, then S_i will tend to be short. Thus this scheme models the case where the server tends to speed up if the queue gets long.

Investigation of such schemes in simple queueing networks is underway; see Jacobs (1977). In particular we know that correlation does affect quantities associated with the queueing networks. Specific analytic results are hard to obtain, but the simplicity of the EARMA models makes it easy to simulate the queues.

4.4. Conclusions

We have presented in this section a number of models for positive valued time series with continuous or discrete ranges which should be useful in modelling the interval or differential count processes of point processes which occur in computer systems. Although the models are not motivated by an underlying physical structure, they have simple probabilistic structure, and therefore should be convenient in modelling and simulating computer systems. Their structural simplicity should also make them easier to fit to data than most standard point process models. In particular, the fact that the specification is in terms of easily measured marginal distributions and second order autocorrelation properties should make rough validation and fitting quite simple. More detailed statistical methods are under development; see Jacobs and Lewis (1977c). Differentiating among related models, for example the three models having exponential marginal distributions and ARMA(1,1) correlation structure, will probably entail use of higher order interval and count spectra.

ACKNOWLEDGMENTS

Professor Peter Lewis gratefully acknowledges support by the Office of Naval Research under Grant NR-42-284 and the National Science Foundation under Grant AF476.

BIBLIOGRAPHY

(1) Anderson, T. W. (1971). Statistical Analysis of Time Series. Wiley, New York.

(2) Bartlett, M. S. (1963). The spectral analysis of point processes. J. Roy. Statist. Soc. B 25, 264-296.

(3) Baskett, F., Chandy, K. M., Muntz, R. R. and Palacios, F, G. (1975). Open, closed, and mixed networks of queues with different clases of jobs. J. ACM 22, 248-260.

(4) Bloomfield, P. (1976). Fourier Analysis of Time Series: An Introduction. Wiley, New York.

(5) Brillinger, D. R. (1972). The spectral analysis of stationary interval functions. Proc. Sixth Berkeley Symp. Math. Statist. and Prob. I, Univ. California Press, Berkeley, 483-513.

(6) Brillinger, D. R. (1975a). The identification of point process systems. Annals Prob. 3, 909-39.

(7) Brillinger, D. R. (1975b). Statistical inference for stationary point processes. In Stochastic Processes and Related Topics, 1. M. L. Puri (ed.), Academic Press, New York, 55-99.

(8) Brown, M. (1972). Statistical analysis of nonhomogeneous Poisson processes. In Stochastic Point Processes, P.A.W. Lewis (ed.), Wiley, New York, 67-89.

(9) Buzen, J. (1971). Queueing network models of multiprogramming. Ph.D. Thesis, Div. of Engineering and Applied Physics. Harvard University, Cambridge, Mass.

(10) Cooley, J.W., Lewis, P.A.W. and Welch, P.D. (1970). The application of the fast Fourier transform algorithm to the estimation of spectra and cross-spectra. J. Sound Vib. 12, 339-352.

(11) Cox, D.R. (1955). Some statistical methods connected with series of events. J. Roy. Statist. Soc. B 17, 129-164.

(12) Cox, D. R. (1963). Some models for series of events. Bull. Int. Statist. Inst. 40, 737-746.

(13) Cox, D.R. (1972). The statistical analysis of dependencies in point processes. In Stochastic Point Processes, P.A.W. Lewis (ed.), Wiley, New York, 55-66.

(14) Cox, D.R. and Lewis, P.A.W. (1966). The Statistical Analysis of Series of Events, Methuen, London; Wiley, New York; Dunod, Paris.

(15) Cox, D.R. and Lewis, P.A.W. (1972). Multivariate point processes. Proc. Sixth Berkeley Symp. Math. Statist. and Prob. III, Univ. California Press, Berkeley, 401-449.

(16) Cox, D.R. and Smith, W.L. (1954). On the superposition of renewal processes. Biometrika 41, 91-99.

(17) Daley, D.J. and Vere-Jones, D. (1972). A summary of the theory of point processes. In Stochastic Point Processes, P.A.W. Lewis (ed.), Wiley, New York, 299-383.

(18) French, A.S. and Holden, A.V. (1971). Alias-free sampling of neuronal spike trains. Kybernetik 9, 165-171.

(19) Gaver, D.P. (1967). Probability models for multiprogrammed computer systems. J. ACM 14, 423-439.

(20) Gaver, D.P. and Lewis, P.A.W. (1977). First order autoregressive Gamma sequences and point processes. To appear.

(21) Gaver, D.P., Lewis, P.A.W. and Shedler, G.S. (1974). Analysis of exception data in a staging hierarchy. IBM J. Res. Devel. 18, 423-435.

(22) Gelenbe, E. and Muntz, R. R. (1976). Probabilistic models of computer systems, Part I (Exact results). Acta Informatica 7, 35-60.

(23) Hannan, E.J. (1970). Multiple Time Series. Wiley, New York.

(24) Haskell, R. and Lewis, P.A.W. (1977). Interval spectra of semi-Markov generated point processes. To appear.

(25) Hawkes, A.G. (1972). Mutually exciting point processes. In Stochastic Point Processes, P.A.W. Lewis (ed.), Wiley, New York, 261-271.

(26) Hawkes, A.G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. J. Appl. Prob. 11, 493-504.

(27) Hunter, D.W. and Shedler, G.S. (1977). Marked multivariate point process models for response times in multiprogrammed systems. IBM Research Report RJ 1963, San Jose, California. To appear in Int. J. Comp. Inf. Sci.

(28) IBM Corp. (1973). Information Management System/360, Version 2 General Information Manual GH20-0765. Armonk, New York.

(29) Jackson, J.R. (1973). Jobshop-like queueing systems. *Manage. Sci.* *10*, 131-142.

(30) Jacobs, P.A. (1977). A closed cyclic queueing network with dependent exponential service times. Submitted for publication.

(31) Jacobs, P.A. and Lewis, P.A.W. (1977a). A mixed autoregressive-moving average exponential sequence and point process (EARMA 1,1), *Adv. Appl. Prob.* *9*, 87-104.

(32) Jacobs, P.A. and Lewis, P.A.W. (1977b). Discrete time series generated by mixtures I: correlational and runs properties. Submitted for publication.

(33) Jacobs, P.A. and Lewis, P.A.W. (1977c). Discrete time series generated by mixtures II: asymptotic properties. Submitted for publication.

(34) Kelly, F.P. (1975). Networks of queues with customers of different types. *J. Appl. Prob.* *12*, 542-554.

(35) Kelly, F.P. (1976). Networks of queues. *Adv. Appl. Prob.* *8*, 416-432.

(36) Lavenberg, S.S. and Shedler, G.S. (1976). Stochastic modeling of processor scheduling with application to data base management systems. *IBM J. Res. Devel.* *20*, 437-448.

(37) Lawrance, A.J. and Lewis, P.A.W. (1975). Properties of the bivariate delayed Poisson process. *J. Appl. Prob.* *12*, 257-268.

(38) Lawrance, A.J. and Lewis, P.A.W. (1977). An exponential moving average sequence and point process (EMA1). *J. Appl. Prob.* *14*, 98-113.

(39) Lewis, P.A.W. (1964a). A branching Poisson process model for the analysis of computer failure patterns. *J. Roy. Statist. Soc. B.* *26*, 1-59.

(40) Lewis, P.A.W. (1964b). Implications of a failure model for the use and maintenance of computers. *J. Appl. Prob.* *1*, 347-368.

(41) Lewis, P.A.W. (1966). A computer program for the statistical analysis of series of events. *IBM Syst. J.* *5*, 202-225.

(42) Lewis, P.A.W. (1967). Non-homogeneous branching Poisson processes. J. Roy. Statist. Soc. B 29, 343-354.

(43) Lewis, P.A.W. (1969). Asymptotic properties and equilibrium conditions for branching Poisson processes. J. Appl. Prob. 6, 355-371.

(44) Lewis, P.A.W. (1970). Remarks on the theory, computation and application of the spectral analysis of series of events. J. Sound Vib. 12, 355-375.

(45) Lewis, P.A.W. (1972). Recent results in the statistical analysis of univariate point processes. In Stochastic Point Processes, P.A.W. Lewis (ed.), Wiley, New York, 1-54.

(46) Lewis, P.A.W. and Robinson, D.W. (1974). Testing for a monotone trend in a modulated renewal process. In Reliability and Biometry. Statistical Analysis of Life length. SIAM, Philadelphia, 163-182.

(47) Lewis, P.A.W. and Shedler, G.S. (1971). A cyclic-queue model of system overhead in multiprogrammed computer systems. J. ACM 18, 119-220.

(48) Lewis, P.A.W. and Shedler, G.S. (1973). Empirically derived micromodels for sequences of page exceptions. IBM J. Res. Devel. 17, 86-100.

(49) Lewis, P.A.W. and Shedler, G.S. (1976). Statistical analysis of non-stationary series of events in a data base system. IBM J. Res. Devel. 20, 465-482.

(50) Mattson, R.L., Gecsei, J., Slutz, D.R. and Traiger, I.L. (1970). Evaluation techniques for storage hierarchies. IBM Syst. J. 9, 78-117.

(51) Moore, C.G., III (1971). Network models for large-scale time-sharing systems. Technical Report No. 71-1, Dept. of Ind. Eng., University of Michigan, Ann Arbor, Michigan.

(52) Mandelbrot, B. (1967). Sporadic random functions and conditional spectral analysis; self similar examples and limits. Proc. Fifth Berkeley Symp. Math. Statist. and Prob. III, Univ. California Press, Berkeley, 155-179.

(53) Oakes, D. (1975). Synchronous and asynchronous distributions for Poisson cluster processes. J. Roy. Statist. Soc. B 37, 238-247.

(54) Oakes, D. (1976). Bivariate Markov processes of intervals. Inf. Control 32, 231-241.

(55) Perkel, D.H., Gerstein, G.L. and Moore, G.P. (1967). Neuronal spike trains and stochastic point processes--II. Simultaneous spike trains. Biophysical J. 7, 419-440.

(56) Reiser, M. and Kobayashi, H. (1976). Queueing networks with multiple closed chains: Theory and computational algorithms. IBM J. Res. Devel. 19, 283-294.

(57) Shedler, G.S. (1977). Response time simulation of multivariate point process models for multiprogrammed jobstreams. IBM Research Report RI 2062, San Jose, California. Submitted for publication.

(58) Slutz, D.R. and Traiger, I.L. (1972). Determination of hit ratios for a class of staging hierarchies. IBM Research Report RJ 1044, San Jose, California.

(59) Synder, D.L. (1976), Random Point Processes, Wiley, New York.

(Keywords: Univariate and Multivariate Series of Events, Statistical Analysis of Point Processes, Spectral Analysis, Trend Analysis, Positive Multivariate Time Series, Computer System Models, Queueing Networks, Stochastic Point Processes; Computer Reliability, Cluster Processes, Page Exception Processes, Nonhomogeneous Poisson Process, Multiprogramming, Multiprogrammed Computer Systems. Autoregressive Sequences, Cluster Process, Discrete Time Series, Self-Similar Processes.)